# Optimistic Active Learning using Mutual Information

Yuhong Guo and Russell Greiner
University of Alberta, Edmonton, Canada

**Active Learning:** A process of sequentially deciding which unlabeled instance to label, with the goal of producing the best classifier with limited number of labelled instances.

**Idea:** an optimistic active learner that exploits the *discriminative partition information* in the unlabeled instances, makes an *optimistic assessment* of each candidate instance, and temporarily *switches to a different policy* if the optimistic assessment is wrong.

## Optimistic Query Selection + Online Adjustment = Mm+M Algorithm

### Optimistic Query Selection

**1. Most uncertain query selection ( MU ):**

$$\underset{i\in U}{\operatorname{argmax}}\, H(Y_i \mid \mathbf{x}_i, L)$$
$$H(Y_i \mid \mathbf{x}_i, L) = -\sum_{y_i} P(y_i \mid \mathbf{x}_i, L)\log P(y_i \mid \mathbf{x}_i, L)$$

**Shortcoming: ignores the unlabeled data !**

**2. Select the query that maximizes its conditional mutual information about the unlabeled data:**

$$\underset{i\in U}{\operatorname{argmax}}\{H(Y_U \mid \mathbf{X}_U, L) - H(Y_U \mid \mathbf{X}_U, L, (\mathbf{x}_i, y_i))\}$$
$$\Updownarrow$$
$$\underset{i\in U}{\operatorname{argmin}}\, H(Y_U \mid \mathbf{X}_U, L, (\mathbf{x}_i, y_i))$$
$$\Updownarrow \quad \text{use parametric model } P(y \mid \mathbf{x}, \boldsymbol{\theta})$$
$$\underset{i\in U}{\operatorname{argmin}} \sum_u H(Y_u \mid \mathbf{x}_u, \boldsymbol{\theta}_{L+(\mathbf{x}_i, y_i)})$$

**Question: How to determine $y_i$ ?**

**Proposals:**

**(a) Take the expectation wrt $Y_i$ ( MCMI[avg] ):**

$$\underset{i}{\operatorname{argmin}} \sum_y P(y \mid \mathbf{x}_i, \boldsymbol{\theta}_L)\, H(Y_u \mid \mathbf{x}_u, \boldsymbol{\theta}_{L+(\mathbf{x}_i, y)})$$

**Shortcoming: aggravates the ambiguity caused by the limited labelled data.**

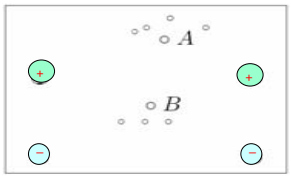
**(b) Take an optimistic strategy: use only the best query label ( MCMI[min] ):**

$$\underset{i\in U}{\operatorname{argmin}}\, f(i)$$
$$\text{where } f(i) = \min_y \sum_u H(Y_u \mid \mathbf{x}_u, \boldsymbol{\theta}_{L+(\mathbf{x}_i, y)})$$

$L$ : set of *labeled* instances
$U$ = index set for *unlabeled* instances
$X_U$ : set of all unlabeled instances

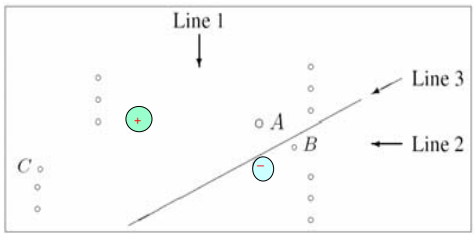
### Online Adjustment

***Optimistic*** **query selection tries to identify the well separated partition. E.g.:**




**Potential problem**: given only a few labelled data points, there might be *many* settings leading to well-separated classes
⟹ our optimistic strategy may "guess" wrong.

**Example**:

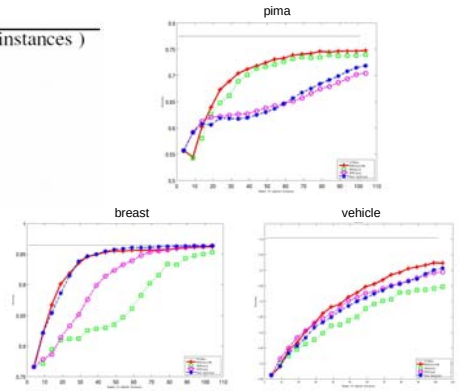



**Online Adjustment** :

- Can easily detect this "guessed wrong" situation, in the immediate next step,
  - Simply compare the actual label for the query with its optimistically predicted label
- Whenever Mm+M guesses wrong,
  - it switches to a different query selection criterion (MU) for the next 1 iteration

### Mm+M Algorithm

MM+M( $U$: indices of unlabeled instances; $L$: labeled instances )
  Repeat
    For each $i \in U$, compute
      $y(i) := \operatorname{argmin}_y \sum_u H(Y_u \mid \mathbf{x}_u, \boldsymbol{\theta}_{L+(\mathbf{x}_i, y)})$
      $f(i) := \sum_u H(Y_u \mid \mathbf{x}_u, \boldsymbol{\theta}_{L+(\mathbf{x}_i, y(i))})$
      *% ie, score based on this minimum $y(i)$ value*
    Let $i^* := \operatorname{argmin}_i f(i)$
      *% instance with optimal* MCMI*[min] score*
    *Purchase true label $w_{i^*}$ for $\mathbf{x}_{i^*}$*
    Remove $i^*$ from $U$; add $(\mathbf{x}_{i^*}, w_{i^*})$ to $L$.
    If $w_{i^*} \neq y(i^*)$, then
      Let $i^* := \operatorname{argmax}_{i\in U} H(Y_i \mid \mathbf{x}_i, L)$
      *Purchase true label $w_{i^*}$ for $\mathbf{x}_{i^*}$*
      Remove $i^*$ from $U$; add $(\mathbf{x}_{i^*}, w_{i^*})$ to $L$.
  until bored.
End MM+M




## Experimental Evaluation

Comparing Mm+M with other Active Learners

| database | vs MU | vs MCMI[min] | vs MCMI[avg] | vs RANDOM | vs MU-SVM |
|---|---|---|---|---|---|
| AUSTRALIAN* | 58/ 0 (+) | 1 / 0 (0) | 24/ 0 (+) | 11 / 14 (−) | 16/ 0 (+) |
| BREAST* | 0 / 0 (−) | 94/ 0 (+) | 56/ 0 (+) | 88/ 0 (+) | 0 / 53 (−) |
| CLEVE* | 56/ 0 (+) | 10/ 0 (+) | 32/ 0 (+) | 26/ 0 (+) | 77/ 0 (+) |
| CORRAL | 5 / 14 (0) | 44/ 2 (+) | 1 / 0 (0) | 13/ 0 (+) | 0 / 23 (−) |
| CRX | 80/ 0 (+) | 0 / 5 (−) | 26/ 0 (+) | 2 / 8 (−) | 13/ 0 (+) |
| DIABETES* | 87/ 0 (+) | 0 / 1 (0) | 87/ 1 (+) | 11 / 12 (+) | 73/ 8 (+) |
| FLARE* | 53/ 0 (+) | 41/ 0 (+) | 55/ 0 (+) | 27/ 2 (+) | 59/ 1 (+) |
| GERMAN* | 42/ 0 (+) | 18/ 0 (+) | 77/ 0 (+) | 0 / 0 (+) | 91/ 0 (+) |
| GLASS2* | 38/ 0 (+) | 4 / 0 (+) | 24/ 0 (+) | 8/ 0 (+) | 17/ 0 (+) |
| HEART* | 30/ 0 (+) | 19/ 0 (+) | 20/ 0 (+) | 0 / 0 (+) | 29/ 0 (+) |
| HEPATITIS* | 7 / 0 (+) | 0 / 0 (0) | 6/ 0 (+) | 0 / 0 (0) | 50/ 0 (+) |
| MOFN | 0 / 33 (−) | 93/ 0 (+) | 67/ 0 (+) | 31/ 6 (0) | 0 / 37 (0) |
| PIMA* | 85/ 2 (+) | 2 / 0 (+) | 84/ 2 (+) | 0 / 0 (0) | 36/ 13 (+) |
| VOTE | 0 / 27 (−) | 97/ 0 (+) | 0 / 0 (−) | 2 / 0 (+) | 0 / 16 (−) |
| IRIS | 68/ 1 (+) | 17/ 0 (+) | 13 / 23 (0) | 38/ 1 (+) | − / − (0) |
| VEHICLE* | 56/ 0 (+) | 84/ 0 (+) | 38/ 0 (+) | 0 / 19 (−) | − / − (0) |
| LYMPHOGRAPHY | 0 / 4 (0) | 0 / 1 (−) | 0 / 30 (−) | 0 / 2 (−) | − / − (0) |
| **TOTAL W/L/T** | **12 / 3 / 2** | **10 / 1 / 6** | **13 / 2 / 2** | **7 / 2 / 8** | **10 / 4 / 0** |
| Signed Rank Test | 12 / 3 / 2 | 12 / 2 / 3 | 13 / 1 / 3 | 10 / 3 / 4 | 9 / 4 / 1 |

Comparing Mm+M vs **MU**, for **PIMA** dataset:
Over *100* sample sizes, **Mm+M** was
• "statistically better" **85** times
• "statistically worse" **2** times
• "tied" **13** times
Signed Rank Test "shows" **Mm+M** is better

Comparing **Mm+M** vs **MCMI[avg]**, over 17 datasets:
**Mm+M** was
• "statistically better" for >5 more sample-sizes: **13** times
• "statistically worse" for >5 more sample-sizes: **2** times
Signed Rank Test "shows" Mm+M is
• better: **13** times
• worse: **1** time

- **Empirical results (over 14+3 databases) show Mm+M works better than**
  - **MU**
  - **MCMI[min]**
  - **MCMI[avg]**
  - **MU (MU-SVM)**
  - **Random**
- **Future work:**
  - **Understand *when* Mm+M is appropriate**
  - **Design further variants**